M. Hildebrandt[1]

Artikel

Algometrisch strafrecht: spiegel of echoput. Kunstmatige intelligentie in het strafrecht[2]

DD 2021/51

Deze bijdrage bespreekt de integratie van kunstmatige intelligentie (KI) systemen in het strafrecht, met name de introductie van software die rechterlijke uitspraken voorspelt, juridische zoekmachines die relevante jurisprudentie, regelgeving en de daarmee verbonden argumentatie aanreikt en zogenaamde risicotaxatie-instrumenten inzake bijvoorbeeld recidive. De kern van deze bijdrage betreft het onderscheid tussen de manier waarop het positieve strafrecht punitief overheidshandelen voorzienbaar maakt en de manier waarop data-gestuurde KI-systemen personen, uitspraken en gedrag voorspelbaar zouden maken. Indien gebruikt om de jurist een spiegel voor te houden kunnen deze systemen wellicht een interessante bijdrage leveren aan de rechtsvinding en de rechtsbedeling. Indien voor waar aangenomen en als objectieve maatstaf overgenomen kan de uitkomst van deze systemen tot ongewenste dynamieken leiden en tot afbreuk van legitieme strafrechtspleging.

1. Introductie

Bij het voorbereidend onderzoek voor deze bijdrage bleken drie concepten koersbepalend: voorzienbaarheid, meetbaarheid en aanstuurbaarheid. Dat laatste is een lelijk woord en misschien ook een lelijk ding. Ik stel juist om die reden voor het nog lelijker overkoepelende neologisme 'algometrisch' te gebruiken.[3] Het refereert tegelijk aan (1) de neiging om overheidsoptreden steeds verder langs algoritmische lijnen te automatiseren en (2) de daartoe doorgevoerde meetbaarheid van menselijk handelen met als oogmerk dat handelen zowel voorzienbaar als aanstuurbaar te maken. Intussen wordt 'handelen' dan wel beschouwd als 'gedrag' zodat het berekenbaar wordt in banale zin, dat wil zeggen omgezet in te aggregeren data die middels dure software kan worden uitgelezen op het niveau van wiskundige patronen. De marketing term voor deze dure software is sinds 1956 kunstmatige intelligentie (KI), waarbij ik graag opmerk dat niet alle aardsvaders van KI content waren met deze verkoopstunt. Herbert Simon zou zich hebben verzet en bleef menen dat het beter 'complexe informatieverwerking' zou heten (volgens zijn collega's zou zo'n saaie term echter geen financiering opleveren en daarmee was de koers gezet).[4] Om misverstanden te voorkomen volg ik hier de onlangs voorgestelde AI Verordening (AIV), die een KI systeem in art. 3(1) als volgt definieert:[5]

---

1 *Onderzoekshoogleraar 'Interfacing Law and Technology' bij de Faculteit Rechten en Criminologie van de Vrije Universiteit Brussel en gewoon hoogleraar 'Smart Environments, Data Protection and the Rule of Law' bij de Faculteit Natuurwetenschappen van de Radboud Universiteit. PI van een ERC AdG inzake computationeel recht, zie www.cohubicol.com en www.journalcrcl.org.*

2 Citeerwijze: M. Hildebrandt, 'Algometrisch strafrecht: spiegel of echoput. Kunstmatige intelligentie in het strafrecht', *DD* 2021/51.

3 In de hoop daarmee liever een to-the-point dan een verhullende term te bezigen, zie hierna de geschiedenis van de term 'kunstmatige intelligentie'.

4 D. Leslie, 'Raging Robots, Hapless Humans: The AI Dystopia', *Nature* 2019 574/7776, p. 32-33.

5 Verordening inzake de Europese aanpak op het gebied van kunstmatige intelligentie, Brussel, 21 april 2021 COM(2021) 206 final 2021/0106 (COD). De Nederlandse vertaling is nog niet beschikbaar, om vertaalverwarring te voorkomen gebruik ik de officiële Engelstalige versie.

"'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with; [...]."

In Annex I worden de betreffende technieken en benaderingen als volgt omschreven:

"(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;
(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;
(c) Statistical approaches, Bayesian estimation, search and optimization methods."

Ik vermoed dat de Europese Commissie het bereik van de term KI-systeem bewust ruim wil houden om zoveel mogelijk bescherming te bieden, vergelijkbaar met het ruime bereik van de term 'persoonsgegeven' in art. 4 van de Algemene Verordening Gegevensbescherming (AVG), die ook door het Hof van Justitie van de EU (HvJEU) ruim wordt uitgelegd om optimale bescherming te kunnen bieden. Ik denk dat dit een verstandig uitgangspunt is, ook waar het gaat om de inzet van KI-systemen binnen het strafrecht. De gedachte is dat met behulp van dit soort systemen de justitiabele zoveel mogelijk kan worden voorzien, teneinde haar mogelijk strafbare gedrag zo vroeg mogelijk in de kiem te smoren (denk aan *crime mapping* of het voorspellen van recidive) en wellicht kan diezelfde justitiabele ook zoveel mogelijk voorzien hoe justitie haar gedrag zal kwalificeren (denk aan het voorspellen van rechterlijke uitspraken). Bovendien kan de advocatuur investeren in datagestuurde innovatie, bijvoorbeeld door het gedrag van individuele strafrechters voorzienbaar te maken (denk aan grote advocatenkantoren die grote bedrijven bedienen die liefst weinig belasting betalen met een zo laag mogelijk risico op strafrechtelijke vervolging of veroordeling).
In deze bijdrage concentreer ik mij op de inzet van KI-systemen in de context van het strafrecht. Dat is iets anders dan de digitalisering van de strafrechtspleging, die ziet op de ontwikkeling van elektronische infrastructuur, zoals elektronische dossiers en virtuele bijeenkomsten met de daarbij vereiste hoge kwaliteit in termen van zowel beveiliging als effectieve gebruiksvriendelijkheid.[6] In plaats van een oppervlakkig overzicht van beide (KI en digitalisering) heb ik gekozen voor wat meer diepgang ten aanzien van KI. Dat neemt niet weg dat – zoals de COVID-19 pandemie heeft aangetoond – het realiseren van een betrouwbare digitale infrastructuur voor de strafrechtspleging cruciaal is, met nadruk op het hoogste niveau van digitale beveiliging en het centraal stellen van de mens als persoon (van opsporingsambtenaar tot verdachte, van hoogste rechter tot griffiemedewerker, van advocaat tot veroordeelde, van officier van justitie tot ICT-ontwikkelaar). Het feit dat de poging om de rechtspraak van zo'n digitale infrastructuur te voorzien in eerste instantie is

6 Zie voor toetsing van deelname aan zitting dan wel onderzoek psychiater via telecommunicatiemiddelen aan de Tijdelijke wet COVID-19 Justitie en Veiligheid bijvoorbeeld: HR 25 september 2020, ECLI:NL:HR:2020:1509, *NJ* 2020/402 m.nt. J. Legemaate; HR 15 december 2020, ECLI:NL:HR:2020:2037, *NJ* 2021/108 m.nt. R.J.B. Schutgens.

mislukt,[7] neemt niet weg dat plannen in die richting serieus moeten worden genomen.[8] Bovendien is de data-gestuurde variant van KI afhankelijk van de digitalisering van de rechtsbedeling, die dan ook tegelijk de doos van Pandora is (of het paard van Troje) ten aanzien van KI. Wanneer regelgeving, rechtspraak en de procesgang eenmaal zijn omgezet in data en metadata zal de verlokking om daarop allerhande KI-systemen te bouwen zich vrijwel onweerstaanbaar doorzetten (misschien moeten we nog eens goed kijken naar Odysseus' list om aan de lokroep van de Sirenen te ontkomen).[9]
Hieronder ga ik eerst kort in op het type voorzienbaarheid dat fundamenteel is voor zowel het recht als het strafrecht, ten minste in een rechtsstaat (2.1). Vervolgens bespreek ik twee wijzen waarop die voorzienbaarheid wordt verdubbeld. Allereerst de uitbreiding van de voorzienbaarheid van overheidshandelen (opsporing, strafoplegging) naar de voorzienbaarheid van de burger (2.2), waarbij ik het betoog toespits op de mogelijkheid om zowel het gedrag van individuele rechters als dat van potentiële verdachten in kaart te brengen en – beweerdelijk – voorspelbaar te maken. Vervolgens bespreek ik de verdubbeling van voorzienbaarheid die aan de orde is als het recht zelf voorspelbaar wordt gemaakt (althans, wanneer we de claims daaromtrent geloven). Daarbij ga ik ervan uit dat de aard van het recht bestaat uit anticipatie op legitiem overheidshandelen (2.1), welke anticipatie wordt vervangen door een voorspelling van rechterlijke uitspraken of relevante juridische argumenten. Met name die juridische argumenten zijn geënt op de anticipatie van hoe de rechter het positieve recht vaststelt, waardoor opnieuw een verdubbeling optreedt, namelijk in de voorspelling van die anticipatie (2.3). In het derde onderdeel bespreek ik het voorspellen van rechterlijke uitspraken (de roborechter) en het voorspellen van relevante juridische bronnen en argumenten (de juridische zoekmachine) (3.1). Daarna ga ik in op risicotaxatieinstrumenten om bijvoorbeeld recidive te voorspellen (3.2). In het laatste onderdeel voorzie ik of en, zo ja, hoe KI-systemen grondrechtenconform geïntegreerd kunnen worden in het strafrecht, met bijzondere aandacht voor enkele bepalingen van de concept AIV (4).

2. De voorzienbaarheid van strafrecht, strafrechter en verdachte en dader

2.1 *De voorzienbaarheid van het strafrecht*

Oliver Wendell Holmes zag het recht als niet meer en niet minder dan de anticipatie op wat de rechter gaat beslissen. Anders dan wat sommigen beweren zei hij niet dat het recht is wat de rechter besluit. Dat zou ons ook niet veel wijzer maken, want het is nu juist wat we proberen te voorzien als we een rechtenstudent, cliënt, ambtenaar, rechtenstudent of gewoon de justitiabele uitleggen wat het recht 'is' in een specifiek geval – voordat de rechter uitspraak heeft gedaan. Deze anticipatie, die eigen is aan het recht, toont het grote belang van de rechtszekerheid, zeker in het strafrecht. De rechtszekerheid eist immers dat we legitieme verwachtingen kunnen ontwikkelen ten aanzien van de gevolgen van ons handelen;

7 *Kamerstukken II* 2017/18, 29279, nr. 420, bijlage brief regering van 13 april 2018 inzake Onderzoek en vervolg programma KEI: Review Board, TRC Consult, 'Definitief Rapport. Quick scan Review KEI. Review op risicobeheersing en basis succescondities voor grote ICT-trajecten', p. 7, waarin de rechterlijke macht o.a. een gebrek aan 'volwassenheid' en 'onvoldoende decisie gerichtheid' wordt verweten, omdat er kennelijk niet voldoende werd meegegaan in het zogenaamde 'agile ontwikkelen'.

8 Zie over de opvolger van KEI:
https://www.rijksoverheid.nl/onderwerpen/rechtspraak-en-geschiloplossing/vernieuwing-in-de-rechtspraak en https://www.rechtspraak.nl/Voor-advocaten-en-juristen/digitalisering-rechtspraak.

9 M. Hildebrandt, *Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology*, Cheltenham: Edward Elgar 2015, p. 156. Over de relatie tussen digitalisering en KI ook: C.M. Aarde & J.E.J. Prins, 'Digitalisering binnen de rechtspraak: van KEI naar Big Data', *Rechtsgeleerd Magazijn Themis* 2016, nr. 2, p. 62–73.

straffen mag alleen voor handelen waarvan men had *moeten* weten dat het strafbaar is. De illusie dat het enkele uitvaardigen van een geschreven regel die zekerheid biedt, miskent de aard van natuurlijke taal, die ambiguïteit herbergt en daarmee verschillende interpretaties mogelijk maakt. Sterker nog, wie meent dat de wet *an sich* zeker stelt welk handelen de facto strafbaar zal zijn, miskent de grondslagen van de rechtsstaat, waar niet de wetgever maar de strafrechter in laatste instantie hoort te beslissen of een handeling strafbaar was.[10] De ambiguïteit van menselijke taal en de grondslagen van de rechtsstaat hangen dan ook samen.[11] Zoals in ander werk uiteengezet, is de rechtsstaat een constructie die mogelijk werd dankzij de proliferatie van het geschreven recht die de multi-interpretabiliteit van het recht vergrootte en daarmee ook de aanvechtbaarheid.[12] Omdat met de opkomst van het geschreven positieve recht de context van uitvaardiging op steeds grotere afstand kwam te staan van de context van toepassing werd de uitleg van het recht steeds minder evident en verscheen uiteindelijk de rechter als onafhankelijke instantie die bij verschil van inzicht de knopen doorhakt. Die rolverdeling was echter niet alleen een kwestie van de coördinatie van een uitdijend regelbestand maar werd tegelijk een manier om willekeur bij de toepassing van het recht door de overheid in te perken. Het positieve recht houdt aldus de spanning in de lucht tussen de voorzienbaarheid en de aanvechtbaarheid van het recht. Het opheffen van die spanning zou leiden tot hetzij een vastgeschroefd recht dat niet buigt naar de omstandigheden van het geval en zich daarmee paradoxaal genoeg leent voor willekeur, dan wel tot de chaos van *Einzelfallgerechtigkeit* die langs andere weg eveneens tot willekeur leidt. Rechtszekerheid veronderstelt om die reden aanvechtbaarheid en heeft in een rechtsstaat, zoals Waldron aangeeft,[13] meer te maken met de argumentatieve aard van het recht dan met een versteende werkelijkheid waarin alles hetzelfde blijft en regels dus probleemloos kunnen worden toegepast.[14] Zo gezien is het belang van de *voorzienbaarheid van de straf* dus niet *dat de justitiable voorzienbaar* en aldus manipulabel wordt omdat de straf zo is gekozen dat de nadelen van bepaald handelen groter worden dan de voordelen (Feuerbach),[15] maar dat die justitiabele wordt aangesproken als een persoon die redenen kan geven voor het eigen handelen en er daarom ook verantwoordelijk voor kan worden gehouden (Hegel).[16] Hoewel Beccaria vaak wordt ingelijfd bij de utilitaristen die de straf zien als een instrument om het geaggregeerde nut van de samenleving te optimaliseren, kan zijn *Over misdaden en straffen* beter worden gelezen als een pleidooi om de persoon van de dader serieus te nemen in het licht van maatschappelijk contract.[17] Zo meende hij dat "vrijheid ophoudt, steeds wanneer de wetten toestaan, dat, in bepaalde gevallen, een mens zou ophouden persoon te zijn, en

10 De praktijk is anders en daarover valt veel te zeggen. In mijn proefschrift heb ik beargumenteerd dat het afdoen buiten geding uiteindelijk minder effectief zal zijn, ook al lijkt het door de uitbreiding van strafbaarheid onvermijdelijk, zie M. Hildebrandt, *Straf(begrip) en procesbeginsel. Een onderzoek naar de betekenis van straf en strafbegrip en naar de waarde van het procesbeginsel* (diss. Rotterdam), Deventer: Kluwer/Sanders Instituut 2002.

11 M. Hildebrandt, 'The Adaptive Nature of Text-Driven Law', *Journal of Cross-Disciplinary Research in Computational Law*, 16 september 2020a.

12 Hildebrandt 2015, hoofdstukken 7 en 8.

13 J. Waldron, 'The Rule of Law and the Importance of Procedure', *Nomos* 2011, 50, p. 3-31.

14 Over de verstening van het recht A.C. 't Hart, *Recht als schild van Perseus*, Arnhem/Antwerpen: Gouda Quint, 1991.

15 Von Feuerbach, *Lehrbuch des gemeinen in Deutschland gültigen Peinlichen Rechts: mit vielen Anmerkungen und Zusatzparagraphen und mit einer vergleichenden Darstellung der Fortbildung des Strafrechts durch die neuen Gesetzgebungen*, Giessen: Georg Friedrich Heyers Verlag 1801; J. M. ten Voorde, 'Strafrechtstheoretische bespiegelingen over afschrikking en generale preventie', *Justitiële verkenningen* 2008, nr. 02; Hildebrandt 2002. Voor de nuance, zie Ten Voorde 2008, p. 59, 61; Hildebrandt 2002, p. 470-473; Feuerbach 1808, p. 13-16.

16 G.W.F. Hegel, *Grundlinien der Philosophie des Rechts*, onder redactie van E. Moldenhauer en K.M. Michel, *Hegel's Werke. Theorie Werkausgabe*, Frankfurt: Suhrkamp 1970, p. 99 en 100.

17 R. Foqué & A.C. 't Hart, *Instrumentaliteit en rechtsbescherming*, Arnhem/Antwerpen: Gouda Quint/Kluwer Rechtswetenschappen 1990.

een ding wordt".[18] Meer Hegel dan Feuerbach, laat staan Bentham;[19] Beccaria's inperking van de straf tot hetgeen nuttig is moet dan ook gelezen worden als een inperking op basis van de noodzakelijkheid in plaats van een poging potentiële daders aan te sturen als waren het pionnen op een schaakbord. Straffen mag alleen als het gerechtvaardigd is door schending van een tevoren uitgevaardigde strafbepaling en alleen voor zover het bijdraagt aan de doelen van de straf, die bij Beccaria niet zien op vergelding in de zin van extra leedtoevoeging maar juist op vermindering van toekomstig leed veroorzaakt door nog te plegen strafbare feiten (in hedendaagse terminologie: generale en specifieke preventie).

2.2 *Op de persoon gespeeld*

Hoe verhoudt deze opvatting van strafrecht en rechtsstaat zich nu tot de dubbele voorzienbaarheid die met een algometrisch strafrecht mogelijk zou worden, namelijk die van de strafrechter enerzijds en die van verdachte en dader anderzijds? Een eerste voorbeeld om de gedachten te bepalen. Sinds 2019 is in Frankrijk een wet van kracht die voorschrijft dat alle rechterlijke uitspraken digitaal openbaar worden gemaakt, mits de privacy van daarin betrokken personen voldoende kan worden beschermd. Tot zover niets nieuws; het is expliciet EU-beleid om alle regelgeving en rechtspraak toegankelijk te maken voor alle burgers op een manier die de tekst ook 'slim' doorzoekbaar maakt.[20] In dat kader werd al in 2019 gesproken van een 'Europese juridische ruimte' die doet denken aan de 'Europese data ruimtes' die de Europese Commissie begin 2020 in haar Europese data-strategie voorstelde.[21] Het gaat er dus niet alleen om bindende regelgeving en rechterlijke uitspraken beschikbaar te stellen aan burgers, maar ook om deze tekst-corpora toegankelijk te maken voor KI-systemen.[22] Kortom: het gaat hier om *recht als data*.[23] Wat de Franse wetgeving bijzonder maakt is dat er een expliciet verbod is bijgevoegd om "de identiteitsgegevens van de rechters te hergebruiken om hun professionele praktijken te evalueren, analyseren, vergelijken of voorspellen".[24] Hierover is van alles te zeggen, maar het geeft aardig aan wat de digitalisering van het recht mogelijk maakt en wat dat zou kunnen betekenen voor de overgang van tekst-gestuurd naar datagestuurd of algometrisch strafrecht. Voor zover het verbod wordt genegeerd – en daar waar zo'n verbod niet bestaat – wordt de strafrechter als individuele beslisser doorzichtig en voorzienbaar gemaakt, en daarmee wordt zij ook bespeelbaar. Niet alleen voor grote

18 D.B. Young, 'Cesare Beccaria: Utilitarian or Retributivist?', *Journal of Criminal Justice* 1983/11, nr. 4, p. 317-26, op p. 321; C. Beccaria, *An Essay on Crimes and Punishments. By the Marquis Beccaria of Milan. With a Commentary by M. de Voltaire. A New Edition Corrected*, Gale ECCO, Print Editions, 2010, p. 79.

19 Hoewel Bentham Beccaria graag prees als ware hij een utilitarist van het zuiverste water, zie Young 1983, p. 318.

20 Zie R.W. Strohmeier, DG Publications Office EU, in: *The contribution of official gazettes to the creation of a European legal space* (Brussel 2019), te raadplegen via: https://op.europa.eu/en/web/forum_official_gazettes/documents-links.

21 Zie inmiddels het Voorstel voor een Verordening van het Europees Parlement en de Raad betreffende Europese datagovernance (Datagovernanceverordening), COM/2020/767 final, 25 november 2020.

22 Ik volg hier de terminologie van het voorstel voor een Verordening inzake de Europese aanpak op het gebied van kunstmatige intelligentie, zie art. 3(1). In Annex III onder 6 en 8 worden KI systemen die worden gebruikt binnen de rechtshandhaving (*law enforcement*) en binnen de rechterlijke macht (voor zover het betreft de rechtstoepassing) gekwalificeerd als 'hoog risico'-systemen, waarop een speciaal regiem van toepassing is.

23 M.A. Livermore & D.N. Rockmore (red.), *Law as data: computation, text, and the future of legal analysis*, Santa Fe: SFI Press 2019.

24 LOI n° 2019-222 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice (1) stipuleert in art. 33.II.1: 'Les données d'identité des magistrats et des membres du greffe ne peuvent faire l'objet d'une réutilisation ayant pour objet ou pour effet d'évaluer, d'analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou supposées.'

advocatenkantoren,[25] maar ook voor de overheid zelf die er belang bij kan hebben te voorzien welk 'beleid' een individuele strafrechter volgens zo'n KI-systeem voert.
Algometrisch strafrecht maakt niet alleen het gedrag van individuele rechters vatbaar voor sturing, het richt zich uiteraard ook op delinquenten, verdachten en potentiële verdachten.[26] Denk bijvoorbeeld aan voorspelling van recidive bij detentiebeslissingen (voorlopige hechtenis, strafoplegging of een vrijheidsbenemende maatregel) op basis van een algometrisch risicotaxatie-instrument, waar ik in paragraaf 3.2 op terug kom.[27] De vooronderstelling is ook hier dat wat meetbaar is kan worden aangestuurd, sterker nog dat is meestal het voor de hand liggende doel van de meetbaarheid. Hoewel in de economische en sociale wetenschappen al geruime tijd bekend is dat wie een maatstaf gebruikt als sturingsmechanisme die maatstaf *als maatstaf* onbruikbaar maakt,[28] is het sturen op gedragsdata bij overheden (en zeker ook in de private sector) in volle gang. Of het nu gaat om de inzet van data-gestuurde technieken om recidive te voorspellen of om de inzet van *natural language processing* (NLP) om rechterlijke beslissingen te voorzien, de toon wordt gezet door een ernstig geloof in data-gestuurde voorspellingen als voorwaarde voor het beïnvloeden van menselijk gedrag.[29] Hoewel dat geloof niet per se tot de gewenste uitkomsten leidt, kan het intussen wel bijdragen aan allerlei vormen van sturing die door de persoon die het betreft niet eenvoudig te voorzien of te weerleggen zijn, bijvoorbeeld omdat de uitkomsten van de meer geavanceerde systemen zich niet zomaar laten doorgronden.

2.3 *De algometrische voorzienbaarheid van het strafrecht*

Een ander voorbeeld van algometrisch strafrecht dient zich aan bij het voorspellen van rechterlijke uitspraken op basis van een data-gestuurde analyse van de inhoud van een relevant corpus van juridische teksten. Het gaat dan niet om het voorzien van individuele personen (rechters dan wel verdachten of daders) maar in zekere zin om het voorspellen van *het recht zelf*. Voor zover het recht zelf een anticipatie is (zie paragraaf 2.1), is opnieuw sprake van een verdubbeling; algometrisch strafrecht verwijst dan immers naar complexe statistieken die voorspellen hoe het recht de beslissing van de rechter zou kunnen voorzien. De manier waarop KI-systemen *voorzien* wijkt intussen af van de manier waarop het recht zelf *voorziet*. KI-systemen danken hun voorspellingen aan de vooronderstelling dat

25 In de VS biedt Westlaw Edge (uiteraard tegen een fikse vergoeding per maand) onder meer de volgende inzichten: 'Get the most relevant highlights for your judge, including ruling tendencies, speed, case type experience, appeals, recent activity, and more. Tailor judge analytics data using filters to narrow your results.' https://legal.thomsonreuters.com/en/products/westlaw/edge/litigation-analytics#judge.

26 Over de schaal waarlangs de redelijke verdenking afglijdt tot een onredelijke schuldpresumptie zie mijn preadvies M. Hildebrandt, 'Data-gestuurde Intelligentie in het Strafrecht', in: *Homo Digitalis*, Handelingen Nederlandse Juristen-Vereniging, Den Haag: Wolters Kluwer 2016, p. 181-196, in het bijzonder p. 185 en noot 94 en p. 191 en verder inzake de verhouding tussen de onschuldpresumptie enerzijds en de redelijke verdenking anderzijds.

27 S.G.C. van Wingerden, L.M. Moerings & J.A. Wilsem, *Recidiverisico en straftoemeting* (Raad voor de rechtspraak, 2011); J. Bijlsma, F. Bex, en G. Meynen, 'Artificiële intelligentie en risicotaxatie', *NJB* 2019, nr. 44, p. 3313-3319. Zie paragraaf 3.2 hieronder.

28 Zie paragraaf 3.2.2 hieronder en D.T. Campbell, 'Assessing the Impact of Planned Social Change', *Evaluation and Program Planning* 1979/ 2, nr. 1, p. 67-90; C. Goodhart, 'Problems of Monetary Management: The UK Experience', in: S. Courakis (red.), *Inflation, Depression, and Economic Policy in the West*, Rowman & Littlefield Publishers 1981, p. 111-44; M. Strathern, '"Improving Ratings": Audit in the British University System', *European Review* 1997/5, nr. 3, p. 305-321.

29 Bijvoorbeeld in de openbare ruimte wordt door het Ministerie van Binnenlandse Zaken volop ingezet op KI-systemen om burgers te 'hypernudgen', zie bijvoorbeeld B. Custers e.a., 'Essaybundel Behoorlijk datagebruik in de openbare ruimte', 2019. In Estonia worden achterstanden bij de rechterlijke macht weggewerkt door KI systemen in te zetten die de slaagkans van zaken doormeten, zie A. Numa, 'Artificial Intelligence as the New Reality of E-Justice', 27 april 2020, https://e-estonia.com/artificial-intelligence-as-the-new-reality-of-e-justice/.

er wiskundige patronen ten grondslag liggen aan juridische teksten (zogenaamde *training data*), die gelijk zijn aan de wiskundige patronen in toekomstige juridische teksten (zogenaamde *test data*). Anders gezegd vooronderstellen deze systemen dat de distributie van historische data gelijk is aan die van toekomstige data. Zonder die vooronderstelling heeft de inzet van KI-systemen geen zin.[30] Het recht zelf is echter een samenspel tussen wetgever, rechter en juridische auteurs, die ieder op eigen wijze bijdragen aan de ontwikkeling van het recht. Het is precies die *rechtsontwikkeling* die in een veranderende samenleving cruciaal is en daarbij een tegenovergestelde vooronderstelling impliceert: de 'distributie van de *training data*' zal als het goed is niet identiek zijn aan die van *test data*.[31] In ander werk sprak ik van het 'bevriezen van de toekomst door het opschalen van het verleden',[32] om aan te geven wat de beperkingen zijn van algometrisch recht. Zolang strafjuristen begrijpen *dat* en *hoe* algometrisch strafrecht onverbiddelijk neigt naar een herhaling van zetten, die in veranderende omstandigheden paradoxaal genoeg tot rechtsonzekerheid zal leiden, kan de inzet van dit soort systemen desondanks nuttig zijn, juist omdat het de strafjurist een spiegel voorhoudt. Rechters zouden bijvoorbeeld in die spiegel kunnen ontdekken dat de ene rechter veel vaker dan de andere asielverzoeken afwijst, dan wel dat zij bepaalde beginselen selectief toepassen zonder dat zij zich daarvan bewust waren. Als het goed is leiden dergelijke inzichten tot reflectie en niet tot klakkeloze aanpassing; zodra we de uitkomst van dit soort systemen als een gegeven gaan beschouwen gaat er iets fundamenteel mis – dan wordt de spiegel een echoput en het recht een filterbubbel.

3. Datagestuurd strafrecht

3.1 *Roborechters en juridische zoekmachines*

3.1.1 *De roborechter*

De idee dat computers in beginsel recht kunnen spreken is al eerder met verve bepleit door bijvoorbeeld Van den Herik.[33] Inmiddels menen Prins en Van der Roest dat:[34]

> "ook indien de rechterlijke macht de hierna door ons geschetste mogelijkheden (nog) niet omarmt, betekent het dat ze slimme data-analyse absoluut zal moeten aanvaarden als een vanzelfsprekende en onvermijdelijke ontwikkeling."

Hoewel de formulering niet zo gelukkig is omdat er een zeker technologisch (of onderliggend economisch) determinisme uit spreekt, doen de auteurs terecht een oproep aan de

30 T. Mitchell, *Machine Learning*, New York: McGraw-Hill Education 1997, p. 6: "We shall see that most current theory of machine learning rests on the crucial assumption that the distribution of training examples is identical to the distribution of test examples. Despite our need to make this assumption in order to obtain theoretical results, it is important to keep in mind that this assumption must often be violated in practice."

31 EHRM 11 februari 2016, appl. nr. 38395/12 (*Dallas v. UK*), r.o. 70 waarin het Hof stelt dat de rechtsontwikkeling (het verfijnen en verhelderen van de betekenis van een strafbaarstelling) noodzakelijk is binnen de juridische traditie in de UK, en dat het strafrechtelijk legaliteitsbeginsel zoals verwoord in art. 6 lid 2 EVRM daaraan niet in de weg staat, zolang verdachte redelijkerwijs kan voorzien of haar gedrag al dan niet strafbaar is en de rechterlijke rechtsvorming consistent is met de essentie van het strafbare feit.

32 M. Hildebrandt, 'Code-Driven Law: Freezing the Future and Scaling the Past', in: C. Markou & S. Deakin (red.), *Is Law Computable? Critical Perspectives on Law and Artificial Intelligence*, Hart Publishing 2020b, vergelijk 't Hart 1991.

33 H.J. van den Herik, *Kunnen computers rechtspreken?* (oratie Leiden), Arnhem: Gouda Quint 1991; M Hildebrandt, 'Oordeelsvorming door mens en machine: heuristieken, algoritmes en legitimatie', in E.T. Feteris e.a. (red.), *Gewogen oordelen. Essays over argumentatie en recht*, Den Haag: Boom juridische uitgevers 2012.

34 J.E.J Prins & J.H.C. van der Roest, 'AI en de rechtspraak', *NJB* 2018/206, p. 206-268.

rechterlijke macht om zich daadwerkelijk te verdiepen in data-gestuurde systemen. Dit is pertinent, juist om te voorkomen dat deze systemen zonder enige kennis van zaken worden geaccepteerd dan wel afgewezen. Prins en Van der Roest bieden vervolgens een kompas aan waarmee de rechterlijke macht koers zou kunnen houden, met name wanneer KI-systemen worden gebruikt bij (1) de rechterlijke oordeelsvorming, (2) het stellen van prioriteiten om bijvoorbeeld de grote toevloed van zaken in goede banen te leiden, en (3) informatievoorziening naar burger en maatschappij over de manier waarop de rechterlijke macht zich van haar taak kwijt.[35] In alle drie de gevallen dreigt de echoput; het recht (ver)valt in herhaling wanneer rechters de rechtsontwikkeling niet langer dienen (met opgaaf van redenen en binnen de bandbreedte van de rechtszekerheid en de rechtvaardigheid). Het is nu vooral zaak een kompas te ontwikkelen dat ons weghoudt van die echoput, zodat we geen nieuw kompas nodig hebben om eruit te klimmen. Naast allerhande praktische aanbevelingen is bij zo'n kompas vooral van belang dat juristen de beperkingen van KI-systemen beter begrijpen.

3.1.2 *Wat doen de relevante KI-systemen?*

De idee van het vervangen of adviseren van rechters door volautomatische computersystemen bestaat al sinds de jaren 80 van de vorige eeuw, grotendeels gebaseerd op juridische expertsystemen. De huidige voorspellingen over 'roborechters' zijn grotendeels gebaseerd op recente ontwikkelingen binnen het onderzoek naar KI-systemen die tekst verwerken *als data*.[36] Data-gestuurde KI-systemen danken hun efficiëntie en snelheid aan de formalisering van instructies. Dat impliceert cruciale beperkingen, bijvoorbeeld ten aanzien van de opgedragen machinaal leesbare 'taak', die de toetssteen vormt die het systeem in staat stelt de eigen prestaties voortdurend te meten, vergelijken en verbeteren. De formulering van die taak vereist een vertaalslag. Zo zal de taak 'uitspraken van het EHRM voorzien die van belang zijn voor de strafrechtspleging' bijvoorbeeld worden geformuleerd als 'het correct voorspellen of het EHRM al dan niet besluit tot schending van artikel 6 of 8 EVRM'. Die laatste formulering is eenvoudiger te formaliseren en functioneert feitelijk als een zogenaamde *proxy* voor de eigenlijke taak. Terwijl de strafrechter wellicht meer belang zal hechten aan de motivering van een rechterlijke uitspraak, omdat juist daarin de daadwerkelijke voorzienbaarheid van het recht zichtbaar wordt, zullen *natural language processing* (NLP) systemen die rechterlijke uitspraken voorspellen zich eerder richten op de uitslag: welke partij heeft gewonnen (OM of verdediging)? Gaat het om een veroordeling of een vrijspraak? Dergelijke 'voorspelsystemen' zijn inmiddels ontwikkeld ten aanzien van de uitspraken van het EHRM, op basis van verschillende type NLP-systemen.[37] Zie bijvoorbeeld de website JURI SAYS waar de voorspellingen van Medvedeva en haar onderzoeksgroep worden vergeleken

35 Prins & Van der Roest 2018 (p. 267) bespreken: kennis over KI-technieken, een afwegingskader om kansen en risico's in kaart te brengen, toetsing aan de kernwaarden van de rechtspraak (onafhankelijkheid, onpartijdigheid, integriteit, transparantie en kwaliteit) en aan de beginselen van behoorlijke rechtspleging en professionele standaarden, met bijzondere aandacht voor de bijdrage van KI aan het gelijkheidsbeginsel en de mogelijke bias die het desondanks kan veroorzaken, voor de privacy van rechters en medewerkers, adequate archivering van gebruikte criteria en algoritmen.

36 Hildebrandt 2012; A-K Oimann, 'AI in de rechterlijke besluitvorming', in: J. de Bruyne & N. Bouteca (red.), *Artificiële intelligentie en maatschappij*, Gompel Svacina 2021, p. 207-222.

37 N. Aletras e.a., 'Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective', *PeerJ Computer Science* 2016/2; I. Chalkidis, I. Androutsopoulos & N. Aletras, 'Neural Legal Judgment Prediction in English', 5 juni 2019, te raadplegeven via: http://arxiv.org/abs/1906.02059; M. Medvedeva, M. Vols & M. Wieling, 'Using Machine Learning to Predict Decisions of the European Court of Human Rights', *Artificial Intelligence and Law*, 2020, 28, nr. 2, p. 237–266; Masha Medvedeva e.a., 'JURI SAYS: An Automatic Judgement Prediction System for the European Court of Human Rights', *Legal Knowledge and Information Systems* 2020, p. 277-280.

met de daadwerkelijke uitkomst.[38] In ander werk heb ik de vooronderstellingen en de implicaties van dit type KI-systemen in detail onderzocht,[39] hier vat ik die bevindingen samen. De belangrijkste vooronderstellingen zijn (1) dat recht kan worden gereduceerd tot tekstcorpora en dat (2) die corpora kunnen worden gereduceerd tot data, dat wil zeggen tot een verzameling van betekenisloze tekstonderdelen ('tokens')[40] met een bepaalde distributie (verdeling binnen de corpora), terwijl (3) in die verzameling relevante wiskundige patronen kunnen worden gedetecteerd. Dat laatste wil zeggen dat een wiskundige functie (model) wordt gezocht die inputdata verbindt met outputdata; bijvoorbeeld een model dat relevante regelgeving, rechtspraak en rechtswetenschappelijke teksten (de inputdata) verbindt met de uitkomst van een geding bij de rechter (de outputdata). Het vinden van een wiskundig model dat steeds de juiste input aan de juiste output koppelt is de heilige graal van machinaal leren (ML) en in dit geval van NLP, of dat nu gebeurt op basis van n-grams, modulaties van BERT of nog meer geavanceerde systemen zoals de 'fantastische' GPT-3.[41]

3.1.3 *Het belang van de rechtsontwikkeling*

Zoals hierboven aangegeven heeft de hele exercitie alleen zin wanneer de distributie van toekomstige juridische data hetzelfde is als die van historische juridische data; het wiskundige model is 'getraind' op bestaande tekstcorpora en heeft dus onvermijdelijk een zogenaamde 'status quo bias'.[42] Omdat bij het voorspellen van rechterlijke uitspraken de nadruk ligt op 'de uitslag' en niet op de onderliggende redenering (daar heeft dit type NLP nu eenmaal geen kaas van gegeten), kan een advocaat of officier van justitie die de software gebruikt, proberen te ontdekken welke beslisregels of factoren doorslaggevend zijn. Dit wordt ook wel *reverse engineering* genoemd. Daarmee kunnen advocaten of officieren van justitie dan hun voordeel doen door een zaak net anders voor te stellen, zodanig dat de kans op een gewenste uitslag toeneemt. Dit is een vorm van *gaming the system*, oftewel het onzichtbaar manipuleren van de uitkomst, niet op basis van juridische argumenten die weerlegd kunnen worden maar op basis van wiskundige slimmigheden. Dit kan bijvoorbeeld door met het betreffende systeem te gaan 'spelen', dat wil zeggen door uit te zoeken welke wiskundige variabelen de uitkomst beïnvloeden in gewenst zin. Dat kunnen voor de hand liggende variabelen zijn, zoals omgevingsfactoren of kenmerken van de verdachte, getuigen of slachtoffer, maar het kunnen ook complexe combinaties van variabelen zijn waar we niet op zouden komen als het KI-systeem er geen patroon in had gezien. De ontwikkelaars van JURI SAYS waarschuwen dan ook tegen de inzet van dit type technologie door bijvoorbeeld de rechterlijke macht, juist omdat het dergelijke manipulatie mogelijk maakt.[43] Zij menen dat dit vooral bij belangrijke zaken moet worden voorkomen. Vanuit het perspectief van de rechtsontwikkeling is het onderscheid tussen belangrijke en minder belangrijke zaken echter niet evident. Eenvoudige, schijnbaar probleemloze zaken kunnen

38 Zie https://www.jurisays.com.

39 M. Hildebrandt, 'Law As Computation in the Era of Artificial Legal Intelligence. Speaking Law to the Power of Statistics', 2017, te raadplegen via: https://papers.ssrn.com/abstract=2983045; M. Hildebrandt, 'Data-driven prediction of judgment. Law's new mode of existence?', 21 oktober 2020, te raadplegen via: https://doi.org/10.31228/osf.io/q5nrm; M. Hildebrandt, 'A philosophy of technology for computational law', 18 november 2020c, te raadplegen via: https://doi.org/10.31228/osf.io/7eykj.

40 Tokens zijn karakters, woorden of subwoorden. Hoewel woorden voor ons betekenis hebben is daar bij een NLP-systeem geen sprake van.

41 Voor een overzicht van de meest geavanceerde NLP-modellen, zie E. Souza Dos Reis e.a., 'Transformers Aftermath: Current Research and Rising Trends', *Communications of the ACM* 64, 2021, nr. 4, p. 154-63.

42 M. Medvedeva, M. Wieling & M. Vols, 'The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems', 18 december 2020, p. 4, te raadplegen via: http://arxiv.org/abs/2012.10301.

43 Medvedeva, Wieling & Vols 2020, p. 4-5 waar zij *reverse engineering* en *adversarial machine learning* bespreken.

leiden tot belangrijke verschuivingen in de rechtsontwikkeling. Bovendien kan een lage boete voor iemand met een laag inkomen een groot verschil maken, en zal het onterecht opleggen van een lage boete niet bijdragen aan de rechtshandhaving noch aan de rechtszekerheid; het is feitelijk een schending van het verbod van willekeur. Die willekeur heeft ook te maken met de manier waarop KI-systemen worden 'verkocht'. Meestal gaat het daarbij om het aanprijzen van de *accuracy* van het systeem. Die *accuracy* kan weliswaar hoog kan zijn, maar dat betekent alleen dat het totaal aantal juiste voorspellingen als percentage van het totaal aantal voorspellingen hoog is. Die maatstaf zegt echter maar weinig over de *precision* of *recall*, die respectievelijk aangeven hoe vaak schending-voorspellingen kloppen als percentage van de werkelijke schendingen en hoe vaak niet-schending voorspellingen kloppen als percentage van de werkelijke niet-schendingen. Juist die laatste twee maatstaven zijn doorslaggevend voor de juistheid van individuele beslissingen. Zo kan het zijn dat bij een *accuracy* van 95% de *precision* desondanks maar 40% is,[44] nu dit afhangt van de verdeling van schending/niet-schending in de *training data*. Alleen al om die reden zou het uitbesteden van uitspraken aan dit type software een hachelijke zaak zijn. Waarbij we dan ook nog moeten opmerken dat de *training data* zelf niet compleet zijn als die alleen gepubliceerde arresten betreft, bijvoorbeeld omdat juist in het voortraject niet-ontvankelijk verklaarde zaken in dat geval buiten zicht blijven.

3.1.4 *Juridische zoekmachines: gemak of het paard van Troje?*

NLP wordt niet alleen gebruikt om rechterlijke uitspraken te voorspellen. Hoewel de idee van de roborechter wellicht tot de verbeelding spreekt, lijkt er vooralsnog weinig animo om delen van het strafrecht door dergelijke software af te laten doen. Intussen worden NLP-toepassingen echter geruisloos ingepast in de bestaande zoeksystemen binnen de oligarchische structuur van de grote juridische uitgevers. In de VS zijn dat vooral LexisNexis en Thomson Reuters (Westlaw and Westlaw Edge),[45] in Nederland Kluwer (die Legal Intelligence opkocht en toevoegde naast de eigen Navigator) en SDU (die Rechtsorde inlijfde en aanbiedt naast de eigen Opmaat).[46] Advocatenkantoren, de rechterlijke macht en justitie kunnen verschillende licenties kiezen, vergelijkbaar met de licenties van universiteitsbibliotheken, waardoor bijvoorbeeld sommige tijdschriften wel en andere niet beschikbaar zijn. Toegang tot de rechtsbronnen (regelgeving en rechtspraak) zou in beginsel voor iedereen vrij moeten zijn, maar in de praktijk is dat maar in beperkte mate het geval, en hoe meer bronnen daadwerkelijk toegankelijk zijn, hoe minder zowel de leek als ook de gespecialiseerde jurist haar weg zal weten te vinden in de hooibergen van potentieel relevante informatie. Daarmee wordt de ontwikkeling van *legal search* een 'logisch' gevolg en een lucratieve business voor degenen die betrouwbare KI-systemen aanbieden op een markt die voor de 'consument' niet te overzien is. Dat laatste is het gevolg van de complexe pakketten die worden aangeboden, die zich niet eenvoudig laten vergelijken met die van de concurrent, maar ook een gevolg van de complexiteit van de betreffende systemen en het onvermogen om de claims die worden gemaakt inzake de functionaliteit te (laten) testen.

44 Een heldere uitleg van deze kwestie: Christian Yates, 'Coronavirus: Surprisingly Big Problems Caused by Small Errors in Testing', *The Conversation* (blog), geraadpleegd 11 juli 2021, http://theconversation.com/coronavirus-surprisingly-big-problems-caused-by-small-errors-in-testing-136700.

45 Zie bijvoorbeeld: https://www.lexisnexis.com/en-us/products/lexis-plus.page en https://legal.thomsonreuters.com/en/products/westlaw (beide bieden 'intelligente' zoekmachines, 'analytics' voor procedures, strategisch advies op basis van data-gestuurde technologie). Er is natuurlijk veel meer, voor een handzaam overzicht: https://sourceforge.net/software/legal-research/ (met een veelheid van filters, ook naar jurisdictie). Voor een commercieel platform dat bemiddelt tussen *legal tech vendors* en potentiële klanten: https://reynencourt.com.

46 Zie het overzicht: https://www.justitia.nl/juridische-portals.

Het vervangen van rechterlijke oordeelsvorming door machinale predicties lijkt vooralsnog niet aan de orde in het strafrecht. Het voorspellen van relevante rechtsbronnen en daaruit afgeleide argumenten is inmiddels dagelijkse praktijk, hoewel ik niet kan overzien in hoeverre dat in Nederland het geval is. Westlaw Edge wordt volgens Thomson Reuters in de VS niet alleen in de *law schools* en grote advocatenkantoren gebruikt maar ook door het departement van justitie, door de federale gerechtshoven en de Supreme Court.[47] Het biedt bijvoorbeeld de mogelijkheid om een *brief* in te voeren teneinde gedetailleerd advies te krijgen inzake relevante jurisprudentie, de juiste versie van geldende regelgeving, inclusief waarschuwingen betreffende gemiste argumentatie.[48] Ook voor deze – reeds ingeburgerde – inzet van KI-systemen gelden de hierboven genoemde beperkingen. Ook hier wordt gewerkt met *training data* en met geformaliseerde taakomschrijvingen; ook hier zegt geclaimde *accuracy* weinig over *precision* of *recall*. Ook hier vindt bij het ontwerp van de software onvermijdelijk een *framing* plaats die zal doorwerken in de uitkomsten.[49] Ook hier zouden andere beslissingen in het ontwerpproces leiden tot andere uitkomsten. Ook hier zal vertrouwen op die uitkomsten leiden tot de nieuwe variant van het Thomas-theorema: "*if machines define a situation as real, it is real in its consequences*."[50]
Gemak dient de mens. De integratie van KI-systemen voegt een nieuwe laag in tussen de jurist en de rechtsbronnen, die echter niet alleen gemak biedt maar ook een spectrum van nieuwe uitdagingen. Waar we als juristen gewend waren aan *close reading* van juridische tekst, zullen we binnenkort wennen aan *distant reading*.[51] Ik hoop dat het openleggen van de wiskundige binnenkant van die nieuwe tussenlaag laat zien dat hier niet alleen efficiencywinst is te verwachten, maar dat het hier ook gaat om het binnenhalen van een paard van Troje.

3.2 *Recidivevoorspellingen*

3.2.1 *Risicotaxatie-instrumenten*

Uit onderzoek van Wingerden et al. van tien jaar geleden leek de toenmalige inzet van risico-inschattingstechnologie geen invloed te hebben op de straftoemeting, behalve wellicht als instrument om de straf te finetunen.[52] Rechters hechtten meer waarde aan de concrete rapportage door de reclassering van de 'criminogene leefgebieden' van de veroordeelde dan aan risico-scores gebaseerd op statistische gemiddelden. Bijlsma et al. beschreven vorig jaar de mogelijke voor- en nadelen van de inzet van dit soort risico-instrumenten, met name ten aanzien van het onderscheid tussen de al genoemde *accuracy*, *precision* en *recall* en de daarmee verbonden trade-offs tussen fout-positieven (waarbij iemand ten onrechte langer gedetineerd wordt of blijft) en fout-negatieven (waarbij slachtoffers zouden kunnen vallen omdat iemand ten onrechte sneller op vrije voeten komt).[53] Zij merken

47 Thomson Reuters, 'Thomson Reuters to Provide Westlaw Edge, Practical Law to US Federal Courts', te raadplegen via: https://www.thomsonreuters.com/en/press-releases/2019/december/thomson-reuters-to-provide-westlaw-edge-practical-law-to-us-federal-courts.html.
48 Zie https://legal.thomsonreuters.com/en/products/westlaw-b.
49 M. Hildebrandt, 'The Issue of Bias. The Framing Powers of Machine Learning', in M. Pelillo & T. Scantamburlo (red.), *Machines We Trust: Perspectives on Dependable AI*, Cambridge, Massachusetts: The MIT Press 2021.
50 Hildebrandt 2015, p. 197 en Robert K. Merton, 'The Self-Fulfilling Prophecy', *The Antioch Review* 1948/8, nr. 2, p. 193. Merton sprak – in navolging van Thomas & Thomas over het feit dat "if men define a situation as real, it is real in its consequences".
51 M Hildebrandt, 'The Meaning and Mining of Legal Texts', in D.M. Berry (red.), *Understanding Digital Humanities: The Computational Turn and New Technology*, London: Palgrave Macmillan, 2011; F. Moretti, *Distant Reading*, London/New York: Verso 2013.
52 Wingerden, Moerings & Wilsem 2011.
53 Bijlsma, Bex & Meynen 2019.

daarbij op dat het mogelijk is om bias in de uitkomst te corrigeren door bijvoorbeeld te eisen dat zwarte en witte verdachten dezelfde kans hebben om ten onrechte als hoog risico te worden aangemerkt (percentage fout-positieven gelijk stellen). Dat kan echter leiden tot meer fout-negatieven, waardoor weliswaar onterechte discriminatie wordt voorkomen maar tegelijk meer ook meer slachtoffers kunnen vallen.[54] Daarnaast wijzen zij op de kwaliteit van de *training data*; als zwarte ingezetenen relatief vaker worden aangehouden dan witte zullen zij oververtegenwoordigd zijn in de data ten opzichte van de daadwerkelijke verdeling van strafbaar relevant gedrag. Dit heeft weer te maken met het feit dat de daadwerkelijke recidive niet bekend is, en wordt gerepresenteerd door bijvoorbeeld een veroordeling, terwijl juist wanneer zwarte mensen vaker worden aangehouden en voorgeleid, die veroordeling geen goede proxy is voor daadwerkelijk recidive. Fogliati et al. laten zelfs zien dat in veel gevallen niet een veroordeling maar een aanhouding als proxy wordt genomen voor recidive, hetgeen de risico-score nog verder vertekent.[55] Belangrijker punt is naar mijn mening dat de onderliggende ontwikkelingen die hebben geleid tot meer strafbaar gedrag in specifieke bevolkingsgroepen onzichtbaar blijven en mogelijk worden versterkt, zeker wanneer politie en justitie op basis van *crime mapping* de aandacht verleggen naar contexten waar algometrische systemen onrust of geweld voorzien.[56]

Hierboven (3.1.2) heb ik aangegeven dat NLP-systemen een machinaal leesbare taakomschrijving behoeven, die formalisering vereist. Die formalisering geldt ook voor de factoren (inputvariabelen) die mogelijk verband houden met de relevante uitkomst (outputvariabele).legde vorig jaar uit dat het in Nederland gebruikte risico-inschattingsinstrument OXREC werkt met de variabele 'deprivation' die wordt bepaald op basis van vijf gegevens: "postcode, ontvangst van bijstandsuitkering (welfare recipiency), werkloosheid, laag opleidingsniveau, criminaliteitscijfers en mediaan inkomen".[57] Het mag duidelijk zijn dat deze gegevens ieder voor zich weer geformaliseerd moeten worden om de variabele meetbaar/machinaal leesbaar te maken. Zoals Van Dijck opmerkt, mogen we ervan uitgaan dat deze gegevens, al dan niet in onderlinge samenhang, correleren met etnische achtergrond: "Indien personen in bepaalde bevolkingsgroepen, met een bepaalde etniciteit, in een bepaalde sociale klasse etc. een grotere kans hebben om te worden opgepakt en te worden veroordeeld dan anderen, zullen genoemde variabelen deze informatie reflecteren."[58] Zolang gegevens over de etnische status van personen niet beschikbaar zijn, kan daarover geen uitsluitsel worden gegeven. Als de huidige versie van art. 10.5 van de voorgestelde AIV kracht van wet zou krijgen,[59] zal dit een juridische grondslag vormen voor de verwerking van bijzondere persoonsgegevens (bijvoorbeeld etniciteit) met als enig doel om bias in KI-systemen te monitoren, detecteren en corrigeren (art. 6.1(c) AVG) en bovendien

54 In dezelfde zin M. Hildebrandt, *Law for Computer Scientists and Other Folk*, Oxford: Oxford University Press 2020, hoofdstuk 11.

55 R. Fogliato e.a., 'On the Validity of Arrest as a Proxy for Offense: Race and the Likelihood of Arrest for Violent Crimes', 11 mei 2021, te raadplegen via: http://arxiv.org/abs/2105.04953.

56 B.E. Harcourt, *Against prediction : profiling, policing, and punishing in an actuarial age*, Chicago: University of Chicago Press 2007.

57 D.W. Braverman e.a., 'OxRec Model for Assessing Risk of Recidivism: Ethics', *The Lancet Psychiatry* 3, 2016, nr. 9, p. 808–809; S. Fazel e.a., 'OxRec model for assessing risk of recidivism: ethics - Authors' reply', *The Lancet Psychiatry* 3, 2016, nr. 9, p. 809–810; S. Fazel e.a., 'Prediction of Violent Reoffending in Prisoners and Individuals on Probation: A Dutch Validation Study (OxRec)', *Scientific Reports* 9, 2019, nr. 1, p. 841; G. van Dijck, 'Algoritmische risicotaxatie van recidive. Over de Oxford Risk of Recidivism tool (OXREC), ongelijke behandeling en discriminatie in strafzaken', *NJB* 2020/95, p. 1784–1790, op p. 1788.

58 Van Dijck 2020, p. 1788.

59 Zie de reeds genoemde Verordening inzake de Europese aanpak op het gebied van kunstmatige intelligentie, Brussel, 21 april 2021 COM(2021) 206 final 2021/0106 (COD).

ook een wettelijke uitzondering rechtvaardigen van het verbod om dergelijke gegevens te verwerken (art. 9.1 en 9.2(g) AVG).
Van Dijck concludeerde dat OXREC – voor zover te toetsen – vaak vals alarm gaf (mensen in een hoge risicogroep plaatste die daar niet horen):[60] "soms zelfs in 70%-80% van de gevallen, wat betekent dat veroordeelden relatief vaak onterecht in een bepaalde risicogroep worden geplaatst. In het meest gunstige geval is er sprake van vals alarm in 40% van de gevallen."[61]
Hij tekent aan dat het niet ondenkbaar is dat menselijke beslissers het nog slechter doen, en benadrukt bovendien dat het de reclassering is die zelfstandig adviseert en de rechter die op basis van kennisname van het advies van de reclassering zelfstandig besluit. Noch de reclassering noch de rechter is gebonden aan enigerlei risico score. Mijns inziens is de kans desondanks groot dat OXREC net als COMPAS in de VS tot discriminatie leidt of gaat leiden, bijvoorbeeld op grond van geslacht of etnische achtergrond. Dit soort instrumenten kosten geld, ze *primen* de beslisser en als het daadwerkelijk zo is dat ze geen invloed hebben gaat het natuurlijk om verspilling van belastinggeld.

3.2.2 *Een maatstaf die als doel wordt gebruikt houdt op een goede maatstaf te zijn*

In de vorige paragraaf (3.1) ging het om rechtsvinding in enge zin: een beslissing over schending van grondrechten of het aanreiken van door het systeem relevant geachte rechtsbronnen en de daarmee samenhangende argumentatielijn. In deze paragraaf gaat het over adviezen van de reclassering en beslissingen van de rechter die beiden een ruime marge hebben waarbinnen zij kunnen adviseren dan wel beslissen. Het rechterlijk oordeel heeft grote gevolgen voor de verdachte of veroordeelde. Het gaat in alle gevallen om de lengte en misschien ook om het type vrijheidsbeneming; ik ga ervan uit dat in het kielzog van de algometrische variant van de forensiche psychiatrie ook de oplegging van TBS zal worden beïnvloed door dit soort risico-inschattingsinstrumenten. De vraag is of, en zo ja, in hoeverre het rechterlijk oordeel inzake vrijheidsbeneming mede moet worden gefaciliteerd door de uitbesteding van risico-inschatting aan algometrische systemen.[62] De schijn van objectiviteit die dergelijke berekeningen aankleeft is gerelateerd aan een hele serie cognitieve bias, bijvoorbeeld *availability bias*, *attribute substitution*, *automation bias*, *default effect*, *confirmation bias*, *time-saving bias*, *zero-risk bias* en *zero-sum bias*. Cognitieve bias is niet noodzakelijkerwijs 'slecht'; zonder dergelijke bias kunnen we überhaupt niet oordelen, en als het goed is dragen ze bij aan een vlotte en betrouwbare inschatting van wat een situatie vereist.[63] Het is echter een illusie te menen dat de inzet van KI-systemen een onwenselijke of verboden bias zouden beperken, of dat ze per se een verbeterde versie aanleveren. Integendeel, er kan sprake zijn van versterking van bestaande onwenselijke bias en van andere typen van bias, die zich bovendien geautomatiseerd doorzetten en vaak lastiger te achterhalen zijn.[64] Daar komt bij dat wanneer de recidive-score wordt gebruikt

60 Zie de Nederlandse web portal van OXREC https://oxrisk.com/oxrec-nl-2-backup/, voor de percentages baseerde Van Dijck zich op Fazel e.a. 2019.
61 Van Dijck 2020, p. 1789.
62 Misschien aardig om de discussie binnen de forensische psychologie en psychatrie er nog eens op na te slaan, onder het motto 'not everything that can be counted counts, and not everthing that counts can be counted', A. Mooij, *De psychische realiteit: psychiatrie als geesteswetenschap*, Amsterdam: Uitgeverij Boom 2006 en idem, *Prudentie en evidentie* (afscheidsrede UU), Den Haag: Boom juridisch 2009. Het beleid is intussen stevig verankerd in meetbare risicotaxatie, zie de brief van de Minister van Rechtsbescherming van 30 juli 2020, Handelingen Tweede Kamer, 33628, nr. 76.
63 F. Schauer, *Profiles Probabilities and Stereotypes*, Cambridge, Massachusetts, London, England: Belknap Press of Harvard University Press 2003; G. Gigerenzer, 'The Bias Bias in Behavioral Economics', *Review of Behavioral Economics* 2018/5, nr. 3-4, p. 303–336.
64 Zie https://www.geckoboard.com/best-practice/statistical-fallacies/.

om mensen mee aan te sturen, deze haar waarde als meetinstrument zal verliezen. Deze 'wetmatigheid' werd in 1975 door de econoom Goodhart geformuleerd als de wet: "that any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes",[65] en in 1997 door Marilyn Strathern kort samengevat als: "if a measure becomes a target, it ceases to be a good measure".[66] Heel eenvoudig vertaald komt die wet erop neer dat de veronderstelling dat de patronen die uit training data worden afgeleid juist wanneer daar beleid op wordt gebaseerd niet zullen overeenkomen met de patronen in toekomstige data. Dat komt volgens Goodhart niet alleen omdat aangepast beleid leidt tot ander gedrag van – in dit geval – potentiële daders, maar ook omdat de beleidmakers en -uitvoerders zich anders gaan gedragen. Niet heel verrassend dunkt mij, maar des te belangrijker om rekening mee te houden bij de aanschaf van dure KI-systemen die dus eigenlijk altijd achter de feiten aanhollen.

4. Grondrechten-conforme integratie van KI-systemen

Ten aanzien van de integratie van KI-systemen in de organisatie van de rechterlijke macht speelt bij het strafrecht in ieder geval de vraag of eventuele algometrische afdoening aan de eisen van art. 6 EVRM zou kunnen voldoen. Daarbij moet worden onderscheiden tussen vervanging en ondersteuning van de rechter, en tussen integratie binnen de zittende magistratuur, dan wel binnen het OM. In het verlengde daarvan zijn ook de integratie binnen de politie en binnen de advocatuur van belang (hier speelt ook de trias; wat betekent het als rechter, bestuur en advocatuur via dezelfde geformaliseerde lens naar het recht gaat kijken). Het antwoord op de vraag of een robo-strafrechter kan voldoen aan art. 6 EVRM lijkt evident. Denk bijvoorbeeld aan het moment waarop de strafrechter in eerste instantie alle uitspraken inzake overtredingen zou uitbesteden aan KI-systemen, met wellicht nog een marginale toets. Zulke systemen zouden aldus een nieuwe laag strafbare feiten geautomatiseerd en grotendeels buiten de rechter om afdoen, naast de reeds aan het OM toegewezen strafbeschikkingen en de aan het bestuursstrafrecht toegewezen sanctionering. De motivering van het antwoord op de vraag hoe zo'n uitbesteding zich verhoudt tot art. 6 EVRM is echter ook van belang voor minder vergaande inzet van vergelijkbare software (zoals bij het voorsorteren van zaken, de vervolgingsbeslissing van het OM, bij beslissingen over detentie,[67] voor de waarde van het bewijs,[68] de relevantie van aangedragen argumenten[69] en gebruik van juridische zoekmachines).

Een strafrechter die zich *laat leiden dan wel vervangen door* een KI-systeem zou, om te beginnen, niet voldoen aan de eis van een tribunaal in de zin van art. 6 EVRM. Bijvoorbeeld omdat geen sprake is van een onafhankelijk (zelfstandig) oordeel door een rechter die voldoet aan de eisen van morele integriteit en juridisch-technische competentie.[70] In de jurisprudentie van het EHRM inzake afdoening van strafzaken door bijvoorbeeld het OM of door bestuursorganen is duidelijk dat zo'n afdoening niet geldt als een uitspraak door een tribunaal. Als dat het geval is, geldt ten aanzien van zo'n uitspraak alsnog het

65 C.A.E. Goodhart, 'Problems of monetary management : the U.K. experience', in: idem, Monetary Theory and Practice: The UK Experience, London: Macmillan Education UK, p. 91-121.
66 Strathern 1997.
67 Van Dijck 2020.
68 B.W. Schermer & J.J. Oerlemans, 'AI, strafrecht en het recht op een eerlijk proces', *Computerrecht* 2020/3.
69 Dit is een onderdeel van 'legal search', zie bijvoorbeeld Nederlandse 'legal intelligence' van Wolters Kluwer https://www.wolterskluwer.com/en/solutions/cheetah en Amerikaanse van Thomson Reuters https://legal.thomsonreuters.com/en/products/westlaw-b.
70 EHRM 1 december 2020, appl. nr. 26374/18 (*Gudmundur Andri Astradsson v. Iceland*), r.o. 219-222.

recht op effectieve toegang tot de rechter, inclusief alle waarborgen van art. 6 EVRM.[71] Dat roept de vraag op wat hiermee is gewonnen en of dit op termijn niet zal leiden tot gebrek aan vertrouwen in de rechter (die zich hier niet meer als rechter gedraagt) en waarschijnlijk tot een verlies aan efficiency dan wel effectiviteit. Wanneer justitiabelen eenvoudig in beroep kunnen gaan tegen de geautomatiseerde beslissing is de kans groot dat hier een verdubbeling van instantie plaatsvindt, waarbij de rechter zich bovendien zal moeten bekwamen in het interpreteren van de uitkomsten van de gebruikte systemen; dat leidt al gauw tot verlies van efficiency. Wanneer beroepsmogelijkheden worden ingeperkt, zoals in bijvoorbeeld het bestuursstrafrecht gebruikelijk is, zal het strafrecht minder effectief worden en om voortdurende monitoring en handhaving vragen, omdat de betreffende norm in mindere mate wordt geïnternaliseerd. De boete op schending van de norm zal worden ervaren als een kostenpost in plaats van een normatieve terechtwijzing, en aldus tot een utilitaire kosten-baten analyse leiden in plaats van integratie van de norm in het eigen handelingsperspectief. Sturen op gedrag in plaats van aanspreken op handelen leidt tot een surveillancesamenleving omdat de enige motivatie om normconform gedrag te vertonen is gelegen in voortdurende monitoring en gedragsmodificatie (het uit de cybernetica bekende reguleringsparadigma dat in Anglo-Amerikaanse discussies over 'regulering' de boventoon voert: '*standard setting*, *monitoring*, *behaviour modification*').[72]

De voorgestelde AIV leert bovendien dat KI-systemen die worden ingezet bij de rechtspraak daarin gelden als hoog-risico systemen (Annex III.8), die moeten voldoen aan een groot aantal eisen, waaronder met name die van menselijk toezicht (art. 14 AIV). Dit toezicht is onder meer gericht op het voorkomen of zoveel mogelijk beperken van het risico dat fundamentele rechten worden geschonden (art. 14.2), zowel wanneer het betreffende KI-systeem voor het beoogde doel wordt gebruikt als wanneer het verkeerd wordt gebruikt (dat vraagt dus adequate anticipatie). 'Redelijkerwijs voorzienbaar verkeerd gebruik' wordt in art. 3(13) gedefinieerd als gebruik voor een ander doel dan bedoeld.[73] Het beoogde doel is van belang omdat dit type systemen alleen getest kunnen worden op basis van een expliciet doel (in de vorm van een machinaal leesbare taak); bij gebruik voor andere doelen kan de geclaimde functionaliteit niet meer worden gegarandeerd en kunnen allerlei neveneffecten optreden die mogelijk impact hebben op de veiligheid en/of de grondrechten, en uiteraard op de kwaliteit en de functionaliteit. Voor zover technisch haalbaar moet de aanbieder van hoog-risico KI-systemen maatregelen voor menselijk toezicht integreren in het ontwerp en de ontwikkeling van het systeem, en anders moet de aanbieder in ieder geval maatregelen voor menselijk toezicht identificeren die de gebruiker van het systeem kan toepassen (14.3).[74] Deze maatregelen moeten de natuurlijke persoon aan wie menselijk toezicht is toevertrouwd onder meer 'in staat stellen' zowel de mogelijkheden als de beperkingen van het systeem 'volledig te begrijpen', 'de werking naar behoren te controleren', te

71 Bijvoorbeeld EHRM 4 maart 2014, appl. nr. 18640/10 (*Grande Stevens v. Italy*), waar een punitieve sanctie was opgelegd door een toezichthouder die niet geldt als tribunaal, met een beroep op de rechter waarvan de onafhankelijkheid en onpartijdigheid niet ter discussie stond. Bij de behandeling van de zaak door die rechter ontbrak echter een openbare zitting, hetgeen alsnog leidde tot schending van art. 6 lid 1 EVRM.

72 Zie daarover bijvoorbeeld M. Hildebrandt, 'Algorithmic regulation and the rule of law', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2018, nr. 2128. Het gaat hier vooral ook over de vraag of meetbaar gedrag een juiste 'proxy' is voor regelgeleid menselijk handelen, zie Hildebrandt 2020.

73 Art. 3(13) AIV: "'reasonably foreseeable misuse' means the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems."

74 'Gebruiker' verwijst hier niet aan de eindgebruiker maar bijvoorbeeld naar een organisatie die het systeem gebruikt om bepaalde diensten aan te bieden (denk aan ziekenhuizen, uitkeringsinstanties, de belastingdienst of advocatenkantoren).

waken tegen 'automation bias', de output van het systeem 'juist te interpreteren', die output eventueel te negeren en ten slotte moet deze natuurlijke persoon in staat zijn 'de werking van het systeem te onderbreken' (14.4).
Met name de eisen dat degene die menselijk toezicht uitoefent het systeem begrijpt en kan afwijken van de uitkomst van het systeem moeten voorkomen dat menselijke tussenkomst een stempel-karakter heeft.[75] Mocht de strafrechter zich bedienen van dit type systemen dan zal de rechterlijke macht zich bij het ontwerpen dan wel inkopen ervan moeten richten op systemen die deze waarborgen omvatten en zich organisatorisch moeten voorbereiden op het daadwerkelijk uitoefenen van menselijk toezicht.
Mijn vraag is of we dit moeten willen; het verplaatst de discussie van die over de inhoud van zaken naar die over de werking van KI-systemen; dit doet denken aan de verplaatsing van de discussie over de juridische rechtvaardiging van beslissingen naar de technische uitleg van hoe geautomatiseerde beslissystemen tot hun uitkomsten komen. Die laatste discussie heeft tot een enorme literatuur geleid rond art. 22 jo. 15 AVG, zowel binnen de informatica als binnen de rechtswetenschappen (over de uitlegbaarheid van KI-systemen en de interpretatie van hun uitkomsten).[76] Hoewel we daar veel van kunnen leren, blijft de vraag of het – gezien bovenstaande – verstandig is de rechtspraak te voorzien van dit type KI-systemen. Het goede nieuws is in ieder geval dat strenge kwaliteitseisen op komst zijn vanwege de AIV, en dat integratie van menselijk toezicht ertoe zal leiden dat rechters die zich bedienen van KI-systemen zich zullen moeten bekwamen in het voldoende begrijpen hoe dergelijke systemen tot hun output komen. Onduidelijk is of het gebruik van KI-systemen door de advocatuur ook als hoog risicosysteem wordt gekwalificeerd, de huidige formulering van Annex III onder 8 AIV is te summier om daar iets over te zeggen. Mij dunkt dat dit zeker het geval zou moeten zijn voor bij de balie ingeschreven advocaten die een eed afleggen waarin zij zich verplichten tot een deontologie die door middel van tuchtrecht wordt gehandhaafd.

5. Afsluiting

Afsluitend concludeer ik dat algometrisch strafrecht ons voor de nodige uitdagingen gaat stellen. Integratie van KI-systemen zal mogelijk in eerste instantie de werkdruk verminderen. Daarna voorzie ik dat het een paard van Troje zal blijken. Eenmaal binnengehaald zal het lastig zijn de geest weer in de fles te krijgen en dat betekent dat gedegen aandacht nodig is voor de wijze waarop en de mate waarin de grondslagen van rechtsstatelijk strafrecht met geïntegreerde KI-systemen kunnen worden omzeild, genegeerd of aangetast. Dat vraagt om daadwerkelijke nieuwsgierigheid naar de mogelijkheden en beperkingen die eigen zijn aan de wiskundige grondslagen van data-gestuurde KI-systemen, en een herschikking van het curriculum van de jurist. Daarin gaat het dan niet om 'leren programmeren voor juristen' of 'knutselen met modulaire software', maar om een grondige kennismaking met de mogelijkheden en beperkingen van een andere discipline (informatica) en een andere professionele praktijk (software engineering), teneinde beter te begrijpen hoe beide zich verhouden tot de grondslagen van het strafrecht en de rechtsstaat.

75 Een vergelijkbare eis wordt gesteld door de EDPB bij de definitie van 'geautomatiseerde beslissingen' in de zin van art. 22 AVG: menselijke tussenkomst door iemand die het systeem niet begrijpt en/of die geen bevoegdheid heeft een andere beslissing te nemen, geldt als een geautomatiseerde beslissing.

76 L. Edwards en M. Veale, 'Slave to the Algorithm? Why a "Right to Explanation" is Probably Not the Remedy You are Looking for', 2017, te raadplegen via: https://papers.ssrn.com/abstract=2972855; Hildebrandt 2018.